Inference from mitochondrial DNA data in forensic identification

Paola Berchialla
Department of Public Health and Microbiology, University of Torino

The interest in maternally inherited mitochondrial DNA (mtDNA) for forensic identification is partially due to the possibility of typing sequences from very small or degradated biological samples.

A general problem in forensic identification arises when a suspect is observed to have a genetic profile (or haplotype) also known to be possessed by the offender whose mtDNA is recovered from a biological sample left at the scene of a crime. An immediate question is how much evidence against the suspect is provided by this matching.

Consider an individual X_1 accused to be the offender. We take a locus as a single position in the mitochondrial DNA sequence. Given that X_1's haplotype has an allele A_i at a locus j, our aim is to compute the probability that another individual X_2, who is not related to X_1, shares the X_1's allele A_i at the same locus j. We refer to this probability as *conditional match probability*. In forensic science the conditional or profile match probability is a natural measure of the weight of evidence in support of the event that the suspect is the offender since it indicates how likely it is another individual shares the suspect's genetic profile.

Calculating this conditional probability needs to take into account the relationship between the known suspect and the unknown person. This relationship may be due to close family membership or to shared evolutionary history.

The current method for evaluating conditional match probabilities is based on the frequency of mtDNA haplotypes within databases. It would be the maximum likelihood estimate of the population proportion if observed sequences were treated as independent, ignoring the genealogical structure. On the other hand complete databases of reference populations have not been yet compiled and many sequences which occur in the population are not represented in the reference sample.

The aim of this paper is to develop a method for analysing data that allows for the effect of the genealogical and mutational history which affect mitochondrial DNA molecule evolution and then computing the match probability in the framework of a fully likelihood based approach.

While most population genetic models for analysing DNA sequence polymorphisms were developed under the infinite-sites model, which assumes that every mutation occurs at a different site in the sequence, for mtDNA this assumption is violated. A more realistic framework is the finite-sites model which allows for multiple substitutions at a single locus.

Conditional on the number of mutations, we consider the mutation process as a random walk between two alleles A_1 and A_2. We assume that A_1 can mutates to A_2 and A_2 to A_1 at a same rate μ, and that alleles can mutate at most once per generation. In this framework, which relies on the assumption of the finite-sites model that not all mutations occuring in the ancestry of a pair of genes lead to observable differences, a result is given for computing the conditional match probability as a function of demographic and mutation parameters.

Finally, we introduce a hierarchical model for allele frequencies that allows for a mutation and a genealogical process based on the standard coalescent model and a coalescent with growth model. This makes it possible to generate observations of mutation and demographic parameters from the post-data distribution in a simulation approach based on the acceptance-rejection method and calculate conditional match probabilities via Monte Carlo method.

1 References

Balding, D. J. and Donnelly, P. (1995), Inference in forensic identification. *J. R. Statist. Soc. A* **158**, 21–53

Foreman, L. A., Smith, A. F. M. and Evett, I. W. (1997) Bayesian analysis of DNA profiling data in forensic identification. *J. R. Statistic. Soc. A*, **160**, 429–469.

Neuhauser, C. (2001). Mathematical models in population genetics, pp. 153–177 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Nicholson G., Smith, A. V., Jonssón, F. Gústafsson, O., Stefánsson, K. & Donnelly, P. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Statist. Soc. A,* **64**, 695–715.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequences data. *Genetics*, **145,** 505–518.

Weir, B. S. (2001) Forensics, pp. 721–739 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Chichester, Wiley.

Wilson, I. J., Weale, M. E & Balding, D. J. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Statist. Soc. A*, **166**, 1–33.